BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
21 November 2002 (21.11.2002)

PCT

(10) International Publication Number
WO 02/093453 A2

(51) International Patent Classification[7]: G06F 19/00

(21) International Application Number: PCT/US02/14665

(22) International Filing Date: 9 May 2002 (09.05.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09/854,183 12 May 2001 (12.05.2001) US

(71) Applicant *(for all designated States except US)*: **X-MINE, INC.** [US/US]; 1000 Marina Blvd., Suite 450, Brisbane, CA 94005 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **HYTOPOULOS, Evangelos** [GR/US]; 605 South Humboldt St., San Mateo, CA 94402 (US). **MILLER, Brett, N.** [CA/US]; 555 Pierce St., Unit 220, Albany, CA 94706 (US). **RAY, Sandip** [IN/US]; 772 20th Avenue, San Francisco, CA 94121 (US).

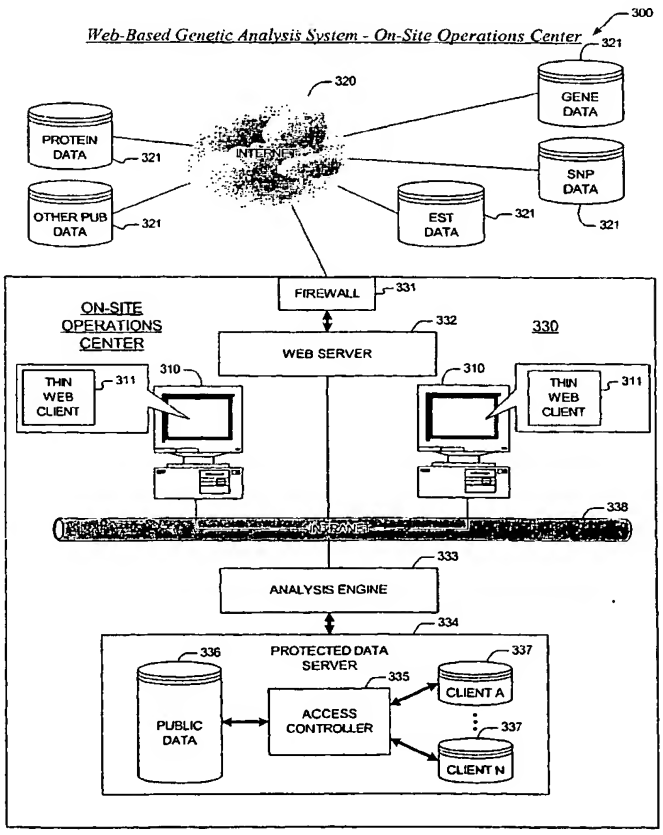
(74) Agent: **HUFFMAN, James, W.**; Huffman Law Group, 1832 N. Cascade Ave., Colorado Springs, CO 80907 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent

[Continued on next page]

(54) Title: WEB-BASED GENETIC RESEARCH ENGINE



(57) Abstract: An apparatus and method are provided for performing genetic analyses in a client-server environment over a data network. The apparatus includes a data server and an analysis engine. The data server stores genetic micro array data sets corresponding to a plurality of users. The analysis engine is coupled to the data server. The analysis engine acquires the genetic micro array data sets for storage, and performs the genetic analyses on the genetic micro array data sets, and provides results of the genetic analyses. The results are provided to corresponding user computers over a data network, where the corresponding user computers employ a thin web client application to configure the genetic analyses and to receive the results.

WO 02/093453 A2

(BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii)) for the following designations AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZM, ZW, ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG)*

— *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii)) for the following designations AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZM, ZW, ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG)*

Published:

— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

TITLE

WEB-BASED GENETIC RESEARCH ENGINE

by

Evangelos Hytopoulos

Brett N. Miller

Sandip Ray

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application is related to the following co-pending U.S. Patent Applications, all of the applications having a common assignee, common inventors, and filed on the same day as this application. The co-pending applications are herein incorporated by reference.

SERIAL NUMBER	DOCKET NUMBER	TITLE
_____	P-2172D1	*ANALYSIS MECHANISM FOR GENETIC DATA*
_____	P-2172D2	*ANALYSIS MECHANISM FOR GENETIC DATA*

BACKGROUND OF THE INVENTION

FIELD OF THE INVENTION

[0002] This invention relates in general to the field of genomics, and more particularly to an apparatus and method for providing web-based genomic analysis capabilities to a plurality of user computers, where the user computers require only a web browser application to configure analyses and to display results of the analyses.

DESCRIPTION OF THE RELATED ART

[0003] Genomics is characterized as a branch of science devoted to investigating and understanding genomes (i.e., a complete set of genes for a given biological organism), where a given genome—in its entirety—is the subject of investigation. It has become increasingly evident that certain biological mutations, diseases, and aberrations result from very complex

and higher-order interrelationships between sets, or clusters, of genes within a genome. Because of these complex interrelationships, those skilled in the art do not restrict their study to subsets of the genome. Rather, it is desirable to analyze the genome as a whole when pursuing a particular path of investigation.

[0004] In prior years, the task of analyzing a genome, even for simple organisms, was deemed infeasible because there was no technique available to easily aggregate the tremendous amounts of data associated with an experiment such that it could be intuitively comprehended by a human being. However, more recent technological advances in the area of gene expression have enabled scientists to progress significantly in the area of genomic analysis, particularly with respect to fostering developments in the biotechnology and pharmaceutical fields.

[0005] Complete genome sequences represent the raw data of genomics. Since the advent of DNA sequencing (1977) approaches, which could readily be incorporated into the standard repertoire of a molecular biology lab, scientists have speculated about the limits to which this technology could be pushed. It was therefore not until as recently as 1995 that the first genome sequence of a free-living organism was published. The organism was Haemophilus influenza, a bacterial species with obvious importance to human health and disease. At 1.83 megabases (Mb) of DNA, complete derivation of the Haemophilus influenza sequence was quite an achievement, and validated the adopted approach of whole genome shotgun sequencing (WGS). In WGS, sequencing is embarked upon without the prior requirement for any genome map upon which cloned genome fragments may be ordered. Rather, the whole genome is fragmented and cloned into vectors to create a library of cloned fragments, which are then end-sequenced in massive quantities and computationally assembled to create a contiguous genomic sequence.

[0006] More recently, Celera™ used the WGS approach to sequence the euchromatin portion of the genome of the fruit-fly, Drosophila melanogaster. Though 120 Mb in length, the real advances made were not in terms of overall complexity of the genome, but more so with methodology. Simply put, the fruit-fly genome was determined in a single year using WGS technology. The chosen strategy placed no reliance upon preexisting maps, but instead extended the capacity of shotgun sequencing to a level, which had previously appeared unattainable.

[0007] We now stand on the verge of knowing the complete sequence of the human genome, which is 3,000 Mbp in length, and consists of approximately thirty thousand genes. The pace at which genomics data acquisition and processing is moving promises more and more genome sequences, including other mammals (including mice), crop plants, and other important pathogens.

[0008] SNP Data

[0009] On average, the genomes of any two human individuals are identical at 99.9 percent of all nucleotides, which translates into one difference for every 1000 bases. While this extremely high degree of identity is striking, the enormous size of the 3,000 Mbp genome means that a 0.1% rate of divergence is still equivalent to over 3 million differences (polymorphisms) between any two people. By far the most common type of polymorphism is a single nucleotide polymorphism (SNP), that is, an alteration of a single base (A,C,G or T) to a different base. SNPs occur (more or less) randomly throughout the genome, which means that about 3 percent of the SNPs will fall within a gene that occupies this same percentage of the genome. SNPs within coding regions of genes (i.e., cSNPs) may be synonymous (i.e., they cause no change in the coding sequence) or nonsynonymous (i.e., they result in amino acid substitution). Most synonymous cSNPs are neutral, resulting in no impact upon health, as is the case for almost all SNPs outside of genes (although important exceptions do occur). Nonsynonymous cSNPs may also be neutral, but many will effect protein expression and/or function, resulting in phenotypes ranging from the benign to serious. Neutral and benign SNPs, while not directly causing disease, have been extremely valuable tools over the years, particularly in the search for elusive disease genes. With the advent of the complete human genome sequence, it is to be expected that most attention will now be focused upon nonsynonymous cSNPs which serve not only as disease markers, but also are frequently found to be the fundamental cause of the disease phenotype. There are approximately 40,000 nonsynonymous cSNPs per person.

[0010] Gene expression Data

[0011] All cells within the human body, with the exception of red blood cells, contain all 23 human chromosomes. And all chromosomes contain all of the genes in the human genome. However, within the chromosomes, not all of the genes are expressed (i.e., turned on) to make

their respective proteins. When genes are expressed, DNA is first transcribed (i.e., copied) to a molecule of messenger RNA. It is this genetic transcription that enables cells to be distinguished from one another as, for instance, nerve cells are distinguished from kidney cells. Accordingly, one of the most prominent areas of genomic investigation today involves the study of gene function through a comparison of the levels of gene expression in body and organ tissues that are in various states of health and disease, and that are also in various states of response to drug therapy.

[0012] Technological advances have provided scientists with a number of platforms that allow the expression levels of thousands of genes to be simultaneously monitored. Presently, the two most widely employed platforms for studying gene expression data are spotted cDNA micro arrays and oligonucleotide micro arrays. Both of these techniques rely upon the deposition of gene sequences upon a substrate in an arrayed pattern. The gene sequences are hybridized with either messenger RNA or cDNA. And by using mixtures labeled with fluorescent dyes, the extent of the hybridization of each gene sequence on the substrate is evaluated by scanning the deposited substrate, or chip, following hybridization. This evaluation involves extrapolating the fluorescent signals from each hybridized gene spot, or cel, into a measure of the abundance of each cDNA and, therefore, messenger RNA populations. These micro array techniques are most powerfully employed when two different messenger RNA populations are labeled with different dyes and compared on the same chip. Under these conditions, the gene expression levels on the chip are expressed as a ratio between conditions of the two messenger RNA (mRNA) populations. Micro array technologies have ushered in a new era of genomic research, primarily because large amount of gene data for one or more experiments can now be assembled for analysis on a single chip.

[0013] Proteomics Data

[0014] Proteomics refers to the study of the complete collection of cellular proteins, in the same way as genomics refers to the complete set of genes. Whereas the genomic sequence data informs us as to which proteins the cell has the potential to make, and microarray expression data provides an approximation of which proteins are made, proteomic approaches define what is happening in a cell in terms of fundamental biochemistry. For instance, putative function can be assigned based upon similarity to other proteins (in the same species

or other species) of known function, or based on mRNA expression patterns, which in most cases serves as a mirror of protein expression patterns.

[0015] Knowing how a protein folds into its final three-dimensional structure is a critical step in understanding function, but it is remarkably difficult to accomplish computationally. Although some secondary structural features such as regions of alpha helix or beta-sheets can be predicted, even the most powerful computing applications cannot yet accurately build a three-dimensional structure based on sequence alone. The only alternative has been structural determination through X-ray crystallography, a time-consuming and extensive procedure which has traditionally been reserved for proteins of special interest. It seems vital to the structural proteomics area of the genomics revolution that either the predictive modeling and/or structural determination of protein structure be further developed to accommodate the high-throughput demand of genomic discovery.

[0016] >It is now contemplated that the human genome comprises upwards of 30,000 genes, approximately 15,000 of which are considered to be potential targets for the development of drug therapies against disease. And even though micro array technologies have allowed the aggregation of thousands of gene expressions on a single chip, manual analysis of this micro array data for the purpose of identifying interesting genes or clusters of genes related to the progression of a certain disease or mutation proves to be an onerous task—chiefly due to the sheer amount of data that is present. Consequently, members of the genomics community have developed a number of *ad hoc* methods for automating certain analytical algorithms that enable scientists and researchers to focus on meaningful gene expressions on a chip rather than all of the data, thus significantly reducing the time that is required to isolate interesting gene sets so that the interesting genes can be investigated further.

[0017] But the current state of the art is limited for several reasons. First, these automated algorithms are typically restricted for application within a certain academic institution or research facility. Second, the automated algorithms are generally developed within the confines of a batch or interpretive programming environment that is designed to enable straightforward and rapid automation of a single complex algorithm in order to solve a particular problem. Examples of such programming environments include Matlab®, MathCad®, and Splus®. All of these programming environments provide researchers with powerful data manipulation and display functions, but programs written for execution in these

interpretive programming environments generally do not port to other platforms, nor can they be easily adapted for application within a client-server environment. Consequently, distribution and use of these automated analysis techniques has been limited to publication in peer journals and the sharing of batch routines between institutional colleagues.

[0018] A scientist at a given research or pharmaceutical development corporation may possess a handful of automated genetic analysis techniques that he/she executes, one-at-a-time, on a particular micro array data set. By evaluating results of each analysis, the scientist identifies interesting genes pertaining to the particular study at hand. The scientist will then execute a number of individual queries over the Internet to a corresponding number of public genetic data repositories to obtain the latest information about the identified interesting genes. For example, the scientist may be interested in various forms of information about the interesting genes to include protein data, single nucleotide polymorphism (SNP) data, and expressed sequence tag (EST) data. To obtain each of these types of data requires that the scientist submit and track a number of queries to different repositories. Furthermore, once information has been provided by all of the repositories, the scientist is then required to aggregate all of the different types of information about each interesting gene into a composite set of information so that a comprehensive evaluation can be made.

[0019] The use of automated techniques for analyzing genetic micro array data has reduced the time that is required to identify interesting genes from years to weeks, but the limited availability of certain techniques, along with the manual effort that is required to determine the most recent information about interesting genes has encumbered scientists. Furthermore, the interpretive programming environment within which analysis algorithms are automated precludes distributed application of a full suite of analytical techniques in today's client-server environment. Researchers are therefore limited in the variety and numbers of analyses that they can perform, and in the manner in which they must interpret results and assimilate information about interesting genes.

[0020] Therefore, what is needed is an apparatus functioning in a client-server environment that allows researchers and scientists to execute a full suite of genetic analyses that enables the scientists to identify the most essential genes in a disease progression pathway and to assess the candidates for their potential as a diagnostic predictor and/or therapeutic target through a panel of other analytics for determining regulatory regions, catalytic domains,

pathway information, subpopulations affected by mutations, whereby the time window for selecting genes from a gene expression data set, or SNP data set, or sequence data, or protein structure is significantly shorter than that which has heretofore been provided.

[0021] In addition, what is needed is a web-based genetic research system that enables a user executing a thin web client on his/her computer to upload gene expression data, SNP data, protein structure data, protein chip data, sequence data, or text (e.g., online journal articles, clinical annotation); to perform a number of different genetic analyses on the data; and to view results of the analyses.

[0022] Furthermore, what is needed is a web-based apparatus that provides aggregated information taken from a number of public data sources about one or more interesting genes, where the interesting genes have been designated by a user through a thin web client interface.

[0023] Moreover, what is needed is a genetic research system that stores micro array data in a common format within a centralized data base, where users remotely access the data base to perform analyses through a web browser application, and where results of the analyses are provided to the users over the Internet.

SUMMARY OF THE INVENTION

[0024] The present invention provides a superior technique that allows researchers and scientists within any Intranet/Internet-enabled institution to access a wide array of integrated automated genetic analytical techniques and result displays. Micro array data in a number of formats can be uploaded over the Intranet/Internet to a centralized data base that converts the data into a common format for storage and processing. After configuring parameters for the genetic analyses via templates provided to a user's web browser, selected analyses are executed within a matter of minutes, and results are provided to the user's web browser, whereby he/she can simultaneously view all results. In addition, the techniques according to the present invention provide the user with an aggregated set of information about designated interesting genes, where the information is obtained from an number of different public or private data sources.

[0025] In one embodiment, an apparatus is provided for performing genetic analyses. The apparatus includes a data server and an analysis engine. The data server stores genetic micro array data sets corresponding to a plurality of users. The analysis engine is coupled to the data server. The analysis engine acquires the genetic micro array data sets for storage, and performs the genetic analyses on the genetic micro array data sets, and provides results of the genetic analyses. The results are provided to corresponding user computers over a data network, where the corresponding user computers employ a thin web client application to configure the genetic analyses and to receive the results.

[0026] One aspect of the present invention features a web-based genetic research system. The web-based research system has a data server, an analysis engine, and a web server. The data server stores micro array data sets corresponding to a plurality of users in a common format, where the micro array data sets are provided to the data server in a variety of formats to include data resulting from cDNA chips (i.e., cDNA format) and data resulting from oligonucleotide chips (i.e., oligonucleotide format). The analysis engine is coupled to the data server. The analysis engine acquires the selected micro array data sets, and performs unsupervised analyses and supervised analyses on the selected micro array data sets, and provides results of the analyses, where the results are provided to a user computer over a data network. The web server is coupled to the analysis engine. The web server transmits and receives transactions over the data network to enable a user executing a web browser on the user computer to configure the analyses and to view the results.

[0027] Another aspect of the present invention contemplates a method for analyzing genetic micro array data sets over a data network via a user computer that is executing a thin web client. The method includes storing the genetic micro array data sets in a common format within a data server; first transmitting/receiving first transactions over the data network to/from the user computer to configure specific analyses to be conducted on specific genetic micro array data sets; within an analysis server, executing the specific analyses, the executing yielding results corresponding to each of the specific analyses; and second transmitting/receiving second transactions over the data network to/from the user computer to provide a user with the results.

BRIEF DESCRIPTION OF THE DRAWINGS

[0028] These and other objects, features, and advantages of the present invention will become better understood with regard to the following description, and accompanying drawings where:

[0029] FIGURE 1 is a diagram illustrating the composition of a typical DNA micro array.

[0030] FIGURE 2 is a flow chart illustrating a method for web-based genetic research according to the present invention.

[0031] FIGURE 3 is a block diagram featuring a web-based genetic analysis system according to the present invention showing configurations for both off-site and on-site operations centers.

[0032] FIGURE 4 is a block diagram showing details of the analysis engine of FIGURE 3.

[0033] FIGURE 5 is a diagram illustrating how a user provides organizational identification within a browser-based login template according to an exemplary embodiment of the present invention.

[0034] FIGURE 6 is a diagram detailing how a user provides project selection information within the browser-based login template.

[0035] FIGURE 7 is a diagram showing features within the login template for selection of a micro array data set.

[0036] FIGURE 8 is a diagram of the login template indicating how the user directs the exemplary embodiment to proceed to configure genetic analyses.

[0037] FIGURE 9 is a diagram of an analysis configuration window according to the exemplary embodiment.

[0038] FIGURE 10 is a diagram of a results template according to the exemplary embodiment that is provided over a data network to a user's thin web client application.

[0039] FIGURE 11 is a diagram of an alternative results template according to the exemplary embodiment that features controls and displays for results corresponding to more than one set of clustered genes.

[0040] FIGURE 12 is a diagram of a quantifier results template according to the exemplary embodiment.

[0041] FIGURE 13 is a diagram featuring a survivor results template according to the exemplary embodiment.

[0042] FIGURE 14 is a block diagram illustrating details of a web server according to the present invention that supports integration of third-party developer applications via enterprise Java Beans.

DETAILED DESCRIPTION

[0043] In light of the above background on the methods and techniques presently employed by scientists and researchers to perform genomic analyses, a detailed description of the present invention will be provided with reference to FIGURES 1 through 13. The present invention overcomes the limitations alluded to above by providing an apparatus and method whereby those within the genomics arts can access and exercise a wide variety of genomic analytical techniques through a simple web browser interface. The analytical techniques are provided in an application programming language that enables deployment in a client-server environment, thus allowing analytical results to be rapidly provided to users.

[0044] Referring to FIGURE 1, a diagram is presented illustrating the composition of a typical DNA micro array 100. The micro array 100 consists of a number of spots 101, or cels 101, that have been placed by deposition in an array of rows 102 and columns 103 on a substrate. Each cel 101 within a particular row 102 represents an expression level of a specific gene 102 across a number of samples 103, or experiments 103, which comprise the columns 103 of the micro array 100. A specific cel 101 represents the expression of a specific gene 102 under a specific experiment 103, as identified by the intersection of the column 103 corresponding to the specific experiment 103 and the row 102 corresponding to the specific gene 102. If the entire human genome were to be represented within the micro

array 100, then roughly 30,000 rows 102 would be present. To date, micro array technologies, such as cDNA and oligonucleotide technologies, allow for representation of up to 15,000 gene rows 102 on a single micro array 100 and for up to 100 experiments 103 per array 100, however technology in this area is progressing so fast that soon chips 100 will be available that can represent the entire genome can across hundreds of experiments 103.

[0045] Operationally, the intensity levels of fluorescent tags within analyzed tissues are utilized to determine a representative expression level for each gene 102 represented on a chip 100. The values of intensity are determined for each cel 101 within the array 100, and these values are stored in an electronic file for manipulation by automated algorithms. In addition to intensity levels that specify expression levels, the electronic file provides an indication of confidence in the reliability of the expression level, a description of the experiment 103, and a name of the gene 102 or gene segment 102 that the cel 101 represents.

[0046] Consider a typical micro array 100 having 75 tissue samples 103 taken from human subjects of different sex, ethnic origin, and additionally having different progressions of various forms of cancer (e.g., breast cancer, prostate cancer, etc.). Furthermore, consider that some of the experiments 103 consist of tissue samples 103 of diseased cells corresponding to various levels of drug therapy. It is the genomic researcher's task to analyze and interpret this huge amount of data in order to isolate one of more genes 102 of interest whose expression correlates to certain phases of a disease, for example. To manually analyze thousands of genes 102 across a number of samples 103 is virtually impossible, even if the expression data 101 is represented graphically.

[0047] Because the data set represented on a micro array 100 is quite large, complex clustering and classification algorithms have been developed to order genes 102 based on their patterns of similar gene expression to enable scientists to rapidly assimilate and interpret the data set. Following execution of many of these algorithms, color coding techniques to graphically portray results. For example, it has become conventional for cDNA data to represent induced genes by the color red and to represent repressed genes by the color green. For oligonucleotide data, various methods are employed to represent analytical results. Many of the algorithms analyze stored intensity levels for genes 102 within the array 100 to identify sets, or clusters, of genes that have statistically similar intensities, thus simplifying the raw data presented within the array 100. Results of these analyses provide scientists with intuitive

as well as integrative interpretations of the raw data. It is beyond the scope of the present application to provide an in-depth discussion of each of the algorithmic techniques that are employed to reduce the amount of data represented within a genetic micro array 100 into meaningful clusters of correlated genes whose properties or impacts can be easily comprehended by a researcher. One skilled in the art will appreciate, however, that such algorithms fall into two major categories: unsupervised analyses and supervised analyses. Unsupervised analytical techniques identify clusters of correlated genes 102 based entirely upon an analysis of cel values 101 within an array 100 (i.e., no outside information is provided for comparison purposes). Supervised analytical techniques evaluate clusters of genes 102 with respect to how their cel values 101 correlate to reference vectors (i.e., information provided from outside sources). Moreover, one skilled in the art will appreciate that it is common practice for a scientist to first execute a series of unsupervised analyses on a micro array data set 100 to provide clusters of genes 102, which are then provided as inputs into supervised analyses. The results of the supervised analyses are next evaluated in order to identity one or more interesting genes 102, i.e., those genes 102 whose expression level may be pivotal with regard to the suppression or progression of a particular disease or mutation. Some of the more common unsupervised analyses include Hierarchical Clustering, Self-Organizing Map (SOM), and Principal Component Analysis (PCA). Examples of supervised clustering techniques include K-Means, Tree Harvesting, and Gene Shaving.

[0048] As alluded to above, virtually all of the supervised and unsupervised analytical techniques have been implemented as batch programs for execution by interpretive programming shells such as Matlab®, Matlab®, and Splus®. These tools are used quite prolifically within the academic and research community because they provide their own set of instructions that enable users to easily prototype very complex algorithms with only a few instructions. These types of programs are widely used because they allow researchers to focus on the algorithm to be implemented rather than the eccentricities of a programming language that may involve other aspects of a computer's architecture. A disadvantage, however, of using these tools to implement genomic analyses is that the analyses execute very slowly in the interpretive environment and, as a result, they cannot be adapted in their current state to a client-server environment. Another disadvantage are that configuration parameters and results for each analysis must be entered and displayed from within the interpretive environment, or imported/exported from/to electronic files. A further disadvantage is that

there is no automated means provided for a scientist or researcher to configure and execute a number of analyses on the same micro array data set 100 and to simultaneously view results of the analyses, thus gaining a level of intuition and insight about the data over that which is available by viewing the results in a sequential manner. And finally, when a researcher does find interesting genes, he/she is forced to employ another platform altogether, like numerous Internet search engines and web-based subscription services, to gather the most recent information about the interesting genes.

[0049] The present invention overcomes the limitations and problems described above by providing an apparatus and method that allows micro array analyses to be performed in an integrated and robust client-server environment. Scientists are not required to execute any of the more prevalently used analysis algorithms on their personal platforms. Rather, a central operations center according to the present invention provides these analyses, along with powerful templates for simultaneously viewing results, such that the analyses and results can be remotely accessed by user clients, whose application requirements extend only to the use of a thin web client, or web browser, such as Microsoft® Internet Explorer® or Netscape® Navigator®.

[0050] Now referring to FIGURE 2, a flow chart 200 is presented illustrating a method for web-based genetic research according to the present invention. Flow begins at block 202 where a user attempts to access a central operations center according to the present invention over a data network through commands provided to a thin web client application on the user's computer. In one embodiment, the data network is the Internet and the thin web client is a web browser that communicates with the central operations center according to TCP/IP protocol. In an alternative embodiment, the data network is an Intranet or local area network and the central operations center is co-located within a client facility. Flow then proceeds to block 204.

[0051] At block 204, the user is required to provided authorization and authentication information to be allowed access to the analyses, and to secure data bases of micro array data that are associated with the user's institution and/or project. In one embodiment, conventional password techniques are employed to enable authorization. In an alternative embodiment, digital certificate/signature techniques are employed to authenticate the user and the user's remote computer. Flow then proceeds to block 206.

[0052] At block 206, the user is allowed to upload micro array data from the remote computer to a data base within the centralized operations center by designating an electronic file location on the user computer by filling in fields of a configuration template provided to the user's web browser. In one embodiment, the data base is employ's Oracle® Data Base Manager. In one embodiment, the electronic file is stored in ASCII format and may contain data resulting from either cDNA chips or oligonucleotide chips. In another embodiment, the configuration template is provided to the user computer via a series of hypertext markup language (HTML) commands. In an alternative embodiment, the template utilizes extensible markup language (XML) commands to transmit the configuration template to the user computer. In another embodiment, Java® applets are provided to the user's web browser to enable control and display of the configuration template.

[0053] The user uploads one or more micro array data sets over the data network to the centralized operations center. Within the central operations center, the data sets are converted from their native protocol into a common format for storage within a data base that is accessible only by authorize users. Flow then proceeds to block 208.

[0054] At block 208, each cel within the micro array data sets is evaluated to determine if its cel value is valid for analysis. One skilled in the art will appreciate that it is common practice within the art to invalidate a particular gene in a given micro array if a significant number of cel values corresponding to the particular gene are invalid for analysis. One skilled in the art will further appreciate that the are number of factors, beyond the scope of this application, that are employed to determine the validity of cel values for a gene to include quality and confidence factors that are provided as part of the raw micro array data set. Following data filtering functions, flow then proceeds to block 210.

[0055] At block 210, the cel values within the micro array data sets are adjusted according to a plurality of normalization techniques to include mean centering, median centering, scale standardization, and linear calibration. Flow then proceeds to block 211.

[0056] At block 211, each gene row is statistically evaluated to determine thresholds of acceptability for gene expression values within each row and certain cel values, otherwise held as invalid, are artificially imputed so that they may be used in ensuing analyses. In one embodiment, raw cel data is imputed if it is found that the raw cel value falls within

statistically acceptable thresholds across all experiments presented on the chip. Flow then proceeds to block 212.

[0057] At block 212, via a configuration template within his/her web browser, the user is allowed to select and configure parameters for an unsupervised analysis technique to be performed on a particular micro array data set. Flow then proceeds to block 214.

[0058] At block 214, the selected unsupervised analysis technique is executed within the central operations center on the micro array data stored within the centralized data base. Flow then proceeds to block 216.

[0059] At block 216, results corresponding to the selected unsupervised analysis technique are stored for subsequent access by the user. Flow then proceeds to block 218.

[0060] At block 218, a plurality of result presentation templates are provided over the data network to the user's web browser to enable the user to simultaneously and selectively view results corresponding to each unsupervised analysis previously performed on the micro array data. Flow then proceeds to block 220.

[0061] At decision block 220, the user is provided with opportunities via a configuration template to designate another unsupervised technique or to continue. If the user selects to perform another unsupervised analysis, then flow proceeds to block 212. If the user elects to continue, then flow proceeds to block 222.

[0062] At block 222, the user is provided with opportunities to select a supervised analysis technique for execution on the micro array data. In one embodiment, the user is allowed to designate cel clusters resulting from previously performed unsupervised analyses as inputs for the supervised technique. In another embodiment, the user is prompted to provide reference vector data for the selected supervised analysis. Flow then proceeds to block 224.

[0063] At block 224, the selected supervised analysis is executed on the micro array data set using inputs and reference vector data as required. Flow then proceeds to block 226.

[0064] At block 226, results of the selected supervised analysis are stored for future display/retrieval by the user. Flow then proceeds to block 228.

[0065] At block 228, a plurality of supervised result presentation templates are provided over the data network to the user's web browser to enable the user to simultaneously and selectively view results corresponding to each supervised analysis previously performed on the micro array data. Flow then proceeds to block 230.

[0066] At decision block 230, the user is prompted to designate another supervised technique or to continue. If the user selects to perform another supervised analysis, then flow proceeds to block 222. If the user elects to continue, then flow proceeds to decision block 232.

[0067] At block 232, based on the user's evaluation of the results of both unsupervised and supervised analyses, one or more interesting genes are designated for further research. In one embodiment, the user is allowed to select these genes directly from result presentation templates. Flow then proceeds to block 234.

[0068] At block 234, the central operations center transmits a plurality of queries over the data network to a plurality of public and/or private data repositories containing different types of information pertaining to the designated interesting genes. In one embodiment, protein, EST, and SNP data bases are queried. As information is provided to the central operations center from each of the queried data repositories, the information is first parsed information categories to include SNP data, protein structure data, protein chip data, gene sequence data, protein-to-protein interaction data, small molecule-to-protein interaction data, and textual information on genes extracted from gene expression analyses. These disparate data types are analyzed and organized such that an aggregated presentation is provided according to each designated interested gene in a format that is easily comprehended by the user. Flow then proceeds to block 236.

[0069] At block 236, the aggregated information about all of the designated interesting genes is provided over the data network to the user as a composite template and/or data file for download. In one embodiment, the user interactively waits for query results to be returned, aggregated, and provided to the web browser of the user computer. In an alternative embodiment, the user may log off from the operations center and is notified by an electronic mail message that the aggregated information is available for download. In a third embodiment, the aggregated information is sent within the electronic mail message. In a

secure embodiment, the electronic mail message is encrypted so that only authorized personnel can view its contents. Flow then proceeds to block 238.

[0070] At block 238, based on an evaluation of information pertaining to the designated interesting genes, and upon analysis results provided to the user, one or more genes are identified as therapeutic targets and/or predictors of disease progression for further development or investigation with regard to the particular project or avenue of investigation enabled that is of interest to the user. Flow then proceeds to block 240.

[0071] At block 240, the method completes.

[0072] The method described with reference to FIGURE 2 is provided as a basis for presenting apparatus and exemplary embodiments according to the present invention that enable researchers and scientists to perform genomic analyses over the Internet or an internal Intranet in a client-server environment. The apparatus and exemplary embodiments according to the present invention are now described with reference to FIGURES 3 through 14.

[0073] Referring to FIGURE 3A, a block diagram is provided featuring a web-based genetic analysis system 300 according to the present invention. The analysis system 300 includes an off-site central operations center 330 that is accessed over a data network 320 by a plurality of off-site computers 310 belonging to a plurality of users or clients. In one embodiment, the data network 320 is the Internet 320 and the off-site computers 310 are executing a Transport Control Protocol (TCP)/Internet Protocol (IP)-based thin web client application 311 such as Microsoft® Internet Explorer® or Netscape® Navigator®. The operations center 330 has a firewall 331 through which data network packets enter and exit. The firewall 331 is coupled to a web server 332. The web server 332 provides front-end web transaction services for an analysis engine 333. The analysis engine 333 is coupled to a protected data server 334. The protected data server 334 comprises an access controller 335 that couples to a public genetic data base 336 and a plurality of user data bases 337. The public data base 236 consists of micro array data sets and related genetic information that can be accessed by all registered and authenticated user computers 310. Each client data base 337 can only be accessed by user computers 310 having access privileges granted by a corresponding client institution or organization.

[0074] In operation, one or more users in an organization maintains a protected data base 337 of micro array data sets. The micro array data sets are uploaded over the data network 320 from files on the user computers 310. The files are provided in varying industry formats including cDNA formats (i.e., data resulting from cDNA chips) and oligonucleotide formats (i.e., data resulting from oligonucleotide chips) and they are converted into a common format for storage within the client data bases 337. The analysis engine 333 controls the timing and sequencing of user activities for uploading micro array data sets, configuring unsupervised and supervised analyses for execution, and transmitting/downloading results of the analyses for display/storage on the client computers 310. In one embodiment, the analysis engine 333 builds Hypertext Markup Language (HTML) web pages for transmittal over the data network 320 to the clients 310. In an alternative embodiment, the analysis engine 333 builds Extensible Markup Language (XML) pages for distribution to the clients 310. In a Java®-based embodiment, the analysis engine 333 builds, processes, and distributes Java applets to the clients 310. Distributing, or providing Java applets to the clients 310 is accomplished by generating data files for the Java applets to read, then generating an HTML page that calls a selected Java applet, and furthermore providing information to the selected Java applet on where to find the information to display.

[0075] The web server 332 receives and issues data network transactions over the data network 320 to affect the distribution of web pages, or templates, and to receive commands and data from the client machines 310.

[0076] Within both the public data base 336 and the secure client data bases 337, micro array experiment information is contained in tables that tracking raw micro array data. Gene and gene segment identification information is contained in a different set of tables that include a broad set of cross-references to all of known representations of this genes and gene segments across the various public databases 321 that have been developed to provide information on genes.

[0077] The analysis engine 333 includes elements (not shown) to filter out genes with unacceptable cel values, to adjust the cel values, and to generate (i.e., impute) cel values. The analysis engine 333 also includes elements that enable researchers to analyze their corresponding micro array data with unsupervised and supervised analysis techniques. Results of analyses are stored within the client data bases 337 in a manner that allows an

exact recreation of every analysis step that is executed by a user. The result data is retrieved from the data bases 337 and provided to result templates that are transmitted to the user computers 310 over the data network 320. Thin web clients 311 within the user computers 310 translate the result templates into visual displays. In a Java-based embodiment, every result template applet contains information, both for display and print, that will allow a researcher to exactly duplicate the analysis steps resulting in that display. Results of each analysis are provided in a separate result template, which can be simultaneously displayed with other result templates on the client computer 310 for result comparison and result integration purposes.

[0078] When a user has determined one or more genes of interest, the result templates enable the user to select them, and to save them as interesting genes, by assigning a set name and supporting comments, all of which information is stored within their respective client data base 337. Via a research template, the user can select the types of further genetic information that is required. Using the cross-reference information stored in the public data base 336, the analysis engine 333 then issues a series of queries to the many public/private databases 321 containing a wide variety of information relating to the types of information that the user wishes to obtain about the interesting genes.

[0079] By doing these searches automatically, dozens of queries for each gene can be run simultaneously, providing a huge gain over a single scientist sitting at a terminal and submitting each query one after the other. When all of the results to these queries have been returned, the results are parsed out of their native format, and reassembled into a report that is much easier to read, and that shows patterns between the genes in the set. The report is stored within the client data base 337 and is also provided to the user computer 310 as a report template. In one embodiment, the report template is interactively provided to the user computer 310. In an alternative embodiment, the report data is stored and an email message is sent via the web server 332 to the user computer 310. Thus, the user may access the report template at a later time. In an attachment-based embodiment, the report data is sent as an encrypted attachment to the email notification message. And as is alluded to above with reference to the execution of analyses, all queries regarding designated interesting genes are tracked and stored in the client data base 337, to allow for duplication of process.

[0080] An alternative embodiment comprehends the acquisition and storage of the above-noted types of information within a private client data base 337. According to the alternative embodiment, scientists and researchers within an organization will access the information types within their own data base 337 rather than issuing public queries. Such an alternative embodiment is enabled via the system configuration shown in FIGURE 3B.

[0081] Referring to FIGURE 3B, a block diagram is provided featuring a web-based genetic analysis system 300 according to the present invention that has an on-site central operations center 330. The on-site central operations center 330 is accessed over an on-site intranet 338 by a plurality of on-site computers 310 belonging to a plurality of users or clients. In one embodiment, the on-site intranet 338 is a local area network 338 executing according to Ethernet protocol. In addition, most frequently accessed information types for interesting genes are acquired by a particular client and stored within their data bases 337.

[0082] Operationally, within the on-site operations center 330, clients 310 access the analysis engine over the intranet 338 rather than via the internet 320. In addition to accessing privately acquired information for interesting genes, requests for data from the public data sources 321 are still routed over the internet 320 through the firewall 331.

[0083] Alternatively, embodiments of the present invention comprehend a combination of both on-site and off-site operations center configurations where users may access the analysis engine 333 and the protected data server 334 via the internet 320 or via a local intranet 338.

[0084] Now referring to FIGURE 4, a block diagram is presented showing details of the analysis engine 400 of FIGURE 3. The analysis engine 400 has a session manager 410 that couples to both the web server (not shown) via bus 401 and the protected data server (not shown) via bus 402. The analysis engine 400 also includes a data acquisition controller 420 coupled to the session manager 410 via bus 411, an unsupervised analysis controller 430 coupled to the session manager 410 via bus 412, a supervised analysis controller 440 coupled to the session manager 410 via bus 413, a results presentation controller 450 coupled to the session manager 410 via bus 414, and a research assistant controller 460 that couples to the session manager 410 via bus 415. In one embodiment, the session manager 410 and controllers 420, 430, 440, 450, 460 are application program modules coded in C/C++ for execution on a Unix-based or Linux-based platform or a Linux box.

[0085] Operationally, the session manager 410 receives user commands and provides user responses to the web server, which manages packetized communications over the data network. To acquire micro array data, the session manager 410 enables the data acquisition controller 420 to direct the acquisition from the user. The data acquisition controller 420 has data format logic 421 that converts the acquired micro array data into a common format for storage in the protected data server, data filter logic 422 that filters out invalid cel data, and a data imputer 423 that imputes data into otherwise invalid cels. In one embodiment, the data imputer 423 imputes cel data based upon statistical calculations performed within a corresponding micro array. The statistical calculations are performed to provide missing cel values in a given gene row. The missing cel values are calculated based upon values in another row that has the closest pattern to the given row. One skilled in the art will appreciate that there are several different statistical techniques that are employed to determine which row is to be used as a model for generating imputed values for the given row.

[0086] To configure and execute unsupervised analyses, the session manager 410 enables the unsupervised analysis controller 430 for interaction with the client. The unsupervised analysis controller 430 provides a plurality of unsupervised configuration templates 431, which are provided to the user for configuration of specific analyses and associated parameters. The unsupervised analysis controller 430 also has a plurality of unsupervised analysis elements 432 that implement a plurality of unsupervised genetic analysis algorithms.

[0087] To configure and execute supervised analyses, the session manager 410 enables the supervised analysis controller 440 for interaction with the client. The supervised analysis controller 440 provides a plurality of supervised configuration templates 441, which are provided to the user for configuration of specific supervised analyses and associated parameters. The supervised analysis controller 440 also has a plurality of supervised analysis elements 442 that implement a plurality of supervised genetic analysis algorithms.

[0088] For presentation of results to the user, the session manager 410 enables the results presentation controller 450. The results presentation controller 450 includes a plurality of result templates 451, each corresponding to one of the unsupervised/supervised techniques 432/442 provided by the unsupervised/supervised analysis controllers 430/440.

[0089] For designation of interesting genes and follow-one queries, the session manager 410 enables the research assistant controller 460. The research assistant controller 460 includes a plurality of research configuration templates 461 that enable the user to designate one or more interesting genes and to prescribe what types of genetic information he/she requires. The research assistant controller 460 also has a composite information generator 462 that aggregates all information received from a plurality of public information sources into a composite report for presentation to the user. The research assistant controller 460 also includes a plurality of research presentation templates 463 that allow presentation of the composite report to the user via the user's thin web client.

[0090] Having now described method and apparatus according to the present invention for providing a web-based genomic research and analysis system, attention is now directed to FIGURES 5 through 13, where a set of exemplary templates, or windows, according to an exemplary embodiment of the present invention will now be discussed.

[0091] Referring to FIGURE 5, a diagram is presented illustrating how a user provides organizational identification within a browser-based login template 500 according to the exemplary embodiment. Once the user's browser is directed to the address of the central operations center, the logic template 500 is sent to the user's web browser. The template 500 shows an organization identification area 501 that has an organization 502 chooser within, a project selection area 503, a data set selection area 504, and an analysis pipeline initiation area 505. The exemplary embodiment disables controls within all areas 503-505 other than the organization identification area 501, thus requiring the user to select his/her organization via the chooser 502.

[0092] Now referring to FIGURE 6, a diagram is presented detailing how a user provides project selection information within a browser-based login template 600. The login template 600 of FIGURE 6 contains elements like those discussed with reference to FIGURE 5, where the hundreds digit is a 6 instead of a 5. After selecting an organization via chooser 602, the user is provided with a project selection chooser 606 within the project selection area 603. Additionally, the template 600 provides a reference name field 607, a project name field 608, a project description field 609, and a create new project button 610, whereby the user can initiate a new project rather than selecting an existing project via chooser 606.

[0093] Following project selection the template 700 of FIGURE 7 is provided to the user's web browser for selection of a micro array data set. The login template 700 of FIGURE 7 contains elements like those discussed with reference to FIGURE 6, where the hundreds digit is a 7 instead of a 6. In addition, within the data set selection area 704, the template 700 now provides a data set chooser 707 that allows the user to select a specific micro array data set for analysis. Rather than selecting an existing data set, the user is also provided with the capability to upload a micro array data set from a client machine via name field 709, file designation field 710, info field 712, chip type chooser 714, a plurality of access control buttons 715, and an upload data set control 716. Instead of entering a data set file name in field 709 and an info file name in field 712, the data set selection area according to the exemplary embodiment also provides data set file designation browse control 711 and info file designation browse control 713. The browse controls 711, 713 enable transmission of directory structure information from the client machine to the central operations center so that the user can select file names for designation. The designated files are then uploaded from the client machine over the data network when the user selects the upload control 716.

[0094] Following specification or upload of a dataset, the template 800 of FIGURE 8 is provided to the user to allow the user to direct the exemplary embodiment to proceed to configure genetic analyses. The login template 800 of FIGURE 8 contains elements like those discussed with reference to FIGURE 7, where the hundreds digit is an 8 instead of a 7. Additionally, the login template 800 of FIGURE 8 enables a go to pipeline control 808 within the analysis pipeline control area 805. By selecting the go to pipeline control 808, the user directs the exemplary embodiment to configure analyses according to the organization, project, and data set information provided by the user as described with reference to FIGURES 5-7.

[0095] Now referring to FIGURE 9, a diagram is presented of an analysis configuration window 900 according to the exemplary embodiment. The analysis configuration window has an input file information area 910 having display fields 911 that reflect the organization, project, and data set information provided via the user as described with reference to FIGURES 5-7. The analysis configuration window 900 also has a data filtering area 920, providing fields 921 whereby the user can prescribe data filtering parameters to include a percent present parameter and an acceptable gene vector standard deviation parameter. The analysis configuration window 900 also has an experiment normalization area 930 providing

selectors 931 to enable mean normalization of micro array data via mean centering, median centering, or scale standardization techniques.

[0096] The analysis configuration window 900 also has a missing data imputation area 940 that provides selectors 941 for the technique used by the exemplary embodiment to impute missing cel data. The selectors 941 allow the user to choose between nearest neighbor imputation or singular value decomposition imputation. An unsupervised analysis configuration area 950 provides the user with a plurality of unsupervised analysis selectors 951, and (if required) corresponding parameter configuration fields 952 to enable the user to prescribe and configure a particular unsupervised analysis. The analysis configuration window 900 includes a view initial analysis area 960 providing a view initial analysis control 961 and a reset defaults control 962. Via the view initial analysis control 961, the user can select to view results of the unsupervised analysis prescribed in the unsupervised analysis configuration area 950. Via the reset defaults control 962, the user can direct the exemplary embodiment to reset unsupervised analysis parameter fields/selectors 952 to their default values.

[0097] Supervised analyses are enabled and configured via a supervised analysis configuration area 970 of the analysis configuration window 900. Like the unsupervised analysis area 950, the supervised analysis area 970 provides the user with a plurality of supervised analysis selectors 971, and (if required) corresponding parameter configuration fields 972 to enable the user to prescribe and configure a particular supervised analysis.

[0098] A results interface area 980 is depicted within the analysis configuration window 900 providing selectors 981, 982 to enable the user to immediately view results (selector 981) of the supervised analysis prescribed within the supervised analysis configuration area 970 or to be notified (selector 982) via an electronic mail message.

[0099] The analysis configuration window 900 additionally has a view supervised analysis area 990 providing a view supervised analysis control 991 and a reset defaults control 992. Via the view supervised analysis control 991, the user can select to view results of the supervised analysis prescribed in the supervised analysis configuration area 970. Via the reset defaults control 992, the user can direct the exemplary embodiment to reset supervised analysis parameter fields/selectors 972 to their default values.

[00100] Now referring to FIGURE 10, a diagram is presented of a results template 1000 according to the exemplary embodiment that is provided over a data network to a user's thin web client application. The results template 1000 provides results of a classifier analysis, an unsupervised analysis technique according to the present invention. The template has a plurality of term controls 1001 that enable the user to selectively view different gene clusters identified via the classifier analysis technique. In addition, the template 1000 provides controls to direct the exemplary embodiment to display results for clusters selected via the terms controls 1001 in normal (i.e., red/green color code) form (control 1002), as cel intensity values 1003, or by cluster category 1004. The results template 1000 also has a results display area 1005 for displaying results according to user selections via controls 1001-1004. For displayed analysis results, the display area 1005 depicts each cel value 1006 within a selected cluster in addition to providing specific gene designations 1010 in a gene designation area 1009 and experiment designations 1008 in an experiment designation field 1007.

[00101] A descriptive results area 1011 of the classifier results template 1000 provides fields 1012 to display aggregate result parameters of the analysis corresponding to the cluster selected via controls 1001. In addition, the template 1000 includes a supplementary presentation area 1013 that graphically depicts supplementary information 1014 associated with the specific analytical technique. In one embodiment, the supplementary information 1014 is an error indicator 1014.

[00102] Now referring to FIGURE 11, a diagram is presented of an alternative results template 1100 according to the exemplary embodiment that features controls and displays for results corresponding to more than one set of clustered genes. The alternative results template 1100 corresponds to results of an unsupervised analysis technique entitled Blade. Like the template 1000 of FIGURE 10, the template 1100 of FIGURE 11 has a plurality of cluster controls 1101 that enable the user to selectively view different gene clusters identified via the blade analysis technique. In addition, the template 1100 provides controls 1102, 1103 to display results for clusters selected via the cluster controls 1101 in normal (i.e., red/green color code) form (control 1102) or as values (control 1103). The results template 1100 also has a results display area 1105 for displaying results according to user selections via controls 1101-1103. For displayed analysis results, the display area 1105 depicts each cel value 1106 within a selected cluster in addition to providing specific gene designations 1110 in a gene

designation area 1109 and experiment designations 1108 in an experiment designation field 1107.

[00103] A descriptive results area 1111 of the blade results template 1100 provides fields 1112 to display aggregate result parameters of the analysis corresponding to the cluster selected via controls 1101. In addition, the template 1100 includes a supplementary presentation area 1113 that depicts supplementary information 1114 associated with the specific analytical technique represented by the template 1100. In one embodiment, the supplementary information 1114 is a graphical representation of the mathematical approach for selecting the cluster size for each cluster of the Blade technique. In this embodiment, a horizontal axis is provided that represents the number of genes in a cluster. A vertical axis is also provided that represents an R^2 score for each cluster of a given size as described in *Gene Shaving: a New Class of Clustering Methods for Expression Arrays,* by T Hastie, R. Tibshirani, M. Eisen, P. Brown, D. Ross, U. Scherf, J. Weinstein, A. Alizadeh, L. Staudt, D. Botstein. A curve 1114 is provided within the supplemental area 1113 entitled "Real Data Score." The curve 1114 shows the R^2 score for each cluster created from the gene data in the microarray. Another curve 1114 entitled "Random Score" is provided that corresponds to the average R^2 score obtained for the same set of genes but with the expression level for each gene permutted randomly. The supplementary information 1114 shown in the supplemental information area 1113 helps the user to visually comprehend the cluster size that the analytical model has chosen.

[00104] FIGURE 12 is a diagram of a quantifier results template 1200 according to the exemplary embodiment. The quantifier results template 1200 corresponds to results of a supervised analysis technique entitled Quantifier. Like the template 1100 of FIGURE 11, the template 1200 of FIGURE 12 has a plurality of term controls 1201 that enable the user to selectively view different gene clusters identified via the quantifier analysis technique. In addition, the template 1200 provides controls 1202, 1203 to display results for clusters selected via the cluster controls 1201 in normal (i.e., red/green color code) form (control 1202) or as values (control 1003). The results template 1200 also has a results display area 1205 for displaying results according to user selections via controls 1201-1203. For displayed analysis results, the display area 1205 depicts each cel value 1206 within a selected cluster in addition to providing specific gene designations 1210 in a gene designation area 1209 and experiment designations 1208 in an experiment designation field 1207.

[00105] A descriptive results area 1211 of the Quantifier results template 1200 provides fields 1212 to display aggregate result parameters of the analysis corresponding to the cluster selected via controls 1201. In addition, the template 1200 includes a supplemental presentation area 1213 that depicts supplemental information 1214 regarding the results of the specific analytical technique that is employed.

[00106] FIGURE 13 is a diagram featuring a survivor results template 1300 according to the exemplary embodiment. The survivor results template 1300 corresponds to results of a supervised analysis technique entitled Survivor. Like the template 1200 of FIGURE 12, the template 1300 of FIGURE 13 has a plurality of term controls 1301 that enable the user to selectively view different gene clusters identified via the survivor analysis technique. In addition, the template 1300 provides controls 1302, 1303 to display results for clusters selected via the cluster controls 1301 in normal (i.e., red/green color code) form (control 1302) or as values (control 1303). The results template 1300 also has a results display area 1305 for displaying results according to user selections via controls 1301-1303. For displayed analysis results, the display area 1305 depicts each cel value 1306 within a selected cluster in addition to providing specific gene designations 1310 in a gene designation area 1309 and experiment designations 1308 in an experiment designation field 1307.

[00107] A descriptive results area 1311 of the Survivor results template 1100 provides fields 1312 to display aggregate result parameters of the analysis corresponding to the cluster selected via controls 1301. In addition, the template 1300 includes a supplementary presentation area 1313 that depicts supplemental information 1314 regarding the results of the specific analytical technique that is employed.

[00108] Now referring to FIGURE 14, a block diagram 1400 is presented illustrating details of a web server 1401 according to the present invention that supports integration of third-party developer applications via enterprise Java Beans 1406-1408. In the embodiment shown in FIGURE 14, a platform structure is provided that enables developers to provide plug-in applications 1406-1408, thus significantly extending the potential of the research system according to the present invention to allow for virtually any needs that a client may decide upon.

[00109] This component or 'plug-in' capability will be provided through the utilization of Enterprise Java Beans (EJB) support within the web server 1401. The Java-enabled server 1401 includes a plurality of Java servlets 1402-1406 to include a registration servlet 1402, a control servlet 1403, a process servlet 1404, and a legacy data base interface servlet 1405. The process servlet 1404 interfaces to a client intranet via bus 1411. The process servlet 1404 processes transactions to/from the intranet. Third-party plug-in applications 1406-1408 are integrated into the research system according to the present invention via enterprise Java beans 1406-1406, examples of which are depicted in the block diagram 1400. Developers can then extend the platform by providing any applications that can be executed via an extended Java bean 1406-1408, to include C/C++ applications, Perl parsers, and Java code. Bus 1410 interfaces the enterprise Java beans 1406-1408 to servlets 1402-1404.

[00110] When the web server 1401 is installed at a client site, a custom servlet 1405 is provided to interface with a legacy client database, if such a database exists. The legacy data base interface servlet 1405 allows for the research system to 1) access a list of the available data sets within the client's database, and 2) retrieve a requested data set for analysis.

[00111] Although the present invention and its objects, features, and advantages have been described in detail, other embodiments are encompassed by the invention as well. For example, the present invention has been particularly characterized as a web-based system whereby clients access a centralized operations center in order to perform optimizations. However, the scope of the present invention is not limited to application within a client-server architecture that employs the Internet or an Intranet as a communication medium. Direct client connection is also provided for by the system according to the present invention.

[00112] In addition, the present invention has been particularly characterized in terms of servers, controllers, and management logic for the analysis of genomic micro array data. These elements of the present invention can also be embodied as application program modules that are executed on a Unix®-based operating system as described or any other operating system, such as Windows NT® that supports HTML, XML, or Java transactions within a client-server environment.

[00113] Those skilled in the art should appreciate that they can readily use the disclosed conception and specific embodiments as a basis for designing or modifying other

structures for carrying out the same purposes of the present invention, and that various changes, substitutions and alterations can be made herein without departing from the spirit and scope of the invention as defined by the appended claims.

[00114] What is claimed is:

CLAIMS

1. An apparatus for performing genetic analyses, comprising:

a data server, configured to store genetic micro array data sets corresponding to a plurality of users; and

an analysis engine, coupled to said data server, configured to acquire said genetic micro array data sets for storage, and configured to perform the genetic analyses on said genetic micro array data sets, and configured to provide results of the genetic analyses, wherein said results are provided to corresponding user computers over a data network, and wherein said corresponding user computers employ a thin web client application to configure the genetic analyses and to receive said results.

2. The apparatus as recited in claim 1, wherein each of said genetic micro array data sets comprises:

a plurality of experiments, each of said plurality of experiments having cel values corresponding to each of a plurality of genes.

3. The apparatus as recited in claim 2, wherein a particular genetic analysis designates a plurality of gene sets within a particular micro array data set, wherein each of said plurality of gene sets comprises one or more of said plurality of genes that have correlated cel values, whereby one of said plurality of users evaluates said plurality of gene sets to identify one or more interesting genes.

4. The apparatus as recited in claim 3, wherein said each of said genetic micro array data sets is stored within said data server in a common format.

5. The apparatus as recited in claim 4, wherein said each of said genetic micro array data sets is uploaded over said data network to said analysis engine from an associated user computer.

6. The apparatus as recited in claim 4, wherein a plurality of said genetic micro array data sets is provided to said analysis engine as data resulting from cDNA chips.

7. The apparatus as recited in claim 4, wherein a plurality of said genetic micro array data sets is provided to said analysis engine as data resulting from oligonucleotide chips.

8. The apparatus as recited in claim 4, further comprising:

a web server, configured to transmit and receive transactions over said data network, wherein said transactions enable said analysis engine to communicate with said corresponding user computers to acquire said genetic micro array data sets and to perform the genetic analyses, and wherein said transactions enable said corresponding user computers to receive said results.

9. The apparatus as recited in claim 8, wherein said data network comprises the Internet, and wherein said thin web client application comprises a web browser.

10. The apparatus as recited in claim 8, wherein said data network comprises an Intranet, and wherein said thin web client application comprises a web browser.

11. The apparatus as recited in claim 10, wherein said analysis engine comprises:

a plurality of configuration templates, configured to enable a particular user to prescribe, from a particular user computer, parameters associated with performing one or more of the genetic analyses on one of said micro array data sets; and

a plurality of result presentation templates, configured to present, to said particular user, analytical results associated with said one or more of the genetic analyses.

12. The apparatus as recited in claim 11, wherein said analysis engine further comprises:

a results presentation controller, configured to provide said plurality of result presentation templates to said particular user, whereby said particular user may simultaneously be presented with said analytical results.

13. The apparatus as recited in claim 12, wherein selected result presentation templates comprise controls and result presentation areas that enable said particular user to selectively view said correlated cel values for said each of said plurality of gene sets.

14. The apparatus as recited in claim 13, wherein said result presentation areas distinguish said correlated cel values by color coding.

15. The apparatus as recited in claim 11, wherein said one or more of the genetic analyses comprise a plurality of first application program modules, said plurality of first application program modules implementing unsupervised genetic analysis techniques selected from a set that includes Hierarchical Clustering, K-Means Clustering, Principal Component Analysis, and Self-Organizing Map, wherein individual outputs of said plurality of first application program modules can be employed as first inputs for said plurality of first application program modules, and wherein said individual outputs can be compared and contrasted for generation of a meta output.

16. The apparatus as recited in claim 15, wherein said one or more genetic analyses further comprise a plurality of second application program modules implementing supervised genetic analysis techniques, wherein said individual outputs can be further employed as second inputs for said plurality of second application program modules.

17. The apparatus as recited in claim 16, wherein said plurality of first application program modules and said plurality of second application program modules are coded in C/C++.

18. The apparatus as recited in claim 11, wherein said plurality of configuration templates and said plurality of result presentation templates are provided as Java® applets over the Internet to said web browser.

19. The apparatus as recited in claim 5, wherein said analysis engine comprises:
a data acquisition controller, configured to convert data from each of said genetic micro array data sets to said common format for storage within said data server, and configured to impute selected cel values into selected micro array data sets.

20. The apparatus as recited in claim 5, wherein said analysis engine comprises:
a research assistant controller, configured to query public and/or private data repositories for disparate information about said one or more interesting genes, and configured to parse, analyze, and aggregate said information for presentation to said one of said plurality of users, wherein said disparate information can include various types of data such as gene expression data, textual data, sequence data, protein chip data, and SNP data.

21. A web-based genetic research system, comprising:
a data server, for storing micro array data sets corresponding to a plurality of users in a common format, wherein said micro array data sets are provided to said data server in a variety of formats to include data resulting from cDNA chips and data resulting from oligonucleotide chips;
an analysis engine, coupled to said data server, for acquiring said selected micro array data sets, and for performing unsupervised analyses and supervised analyses on said selected micro array data sets, and for providing results of said analyses, wherein said results are provided to a user computer over a data network; and
a web server, coupled to said analysis engine, for transmitting and receiving transactions over said data network to enable a user executing a web browser on said user computer to configure said analyses and to view said results.

22. The web-based genetic research system as recited in claim 21, wherein said analysis engine acquires said selected micro array data sets by uploading associated electronic files over said data network from said user computer.

23. The web-based genetic research system as recited in claim 21, wherein said data network comprises the Internet.

24. The web-based genetic research system as recited in claim 21, wherein said data network comprises an Intranet.

25. The web-based genetic research system as recited in claim 21, wherein said analysis engine comprises:

a plurality of configuration templates, for enabling said user to configure said analyses; and

a plurality of result presentation templates, for enabling said user to view said results.

26. The web-based genetic research system as recited in claim 25, wherein said analysis engine further comprises:

a results presentation controller, for providing said plurality of result presentation templates to said user, wherein, when said results correspond to more than one of said analyses, said user may simultaneously view each of said result presentation templates that corresponds to each of said more than one of said analyses.

27. The web-based genetic research system as recited in claim 26, wherein first result presentation templates comprise controls and result presentation areas that allow said user to selectively view correlated cel values for each of a plurality of gene sets, wherein said each of a plurality of gene sets is identified by one of said analyses.

28. The web-based genetic research system as recited in claim 27, wherein said result presentation areas distinguish said correlated cel values by color coding.

29. The web-based genetic research system as recited in claim 21, wherein said unsupervised analyses comprise a plurality of first application program modules, said plurality of first application program modules implementing unsupervised genetic analysis techniques selected from a set that includes Hierarchical Clustering, K-Means Clustering, Principal Component Analysis, and Self-Organizing Map.

30. The web-based genetic research system as recited in claim 29, wherein said supervised analyses comprise a plurality of second application program modules implementing supervised genetic analysis techniques.

31. The web-based genetic research system as recited in claim 30, wherein said plurality of first application program modules and said plurality of second application program modules are coded in C/C++.

32. The web-based genetic research system as recited in claim 25, wherein said plurality of configuration templates and said plurality of result presentation templates are provided as Java® applets over said data network to said web browser.

33. The web-based genetic research system as recited in claim 21, wherein said analysis engine comprises:

a data acquisition controller, for converting data from said selected micro array data sets to said common format for storage within said data server, and for imputing specific cel values into specific micro array data sets.

34. The web-based genetic research system as recited in claim 21, wherein said analysis engine comprises:

a research assistant controller, for querying public/private data sources for information about interesting genes, said interesting genes being designated by said user, and for aggregating said information for presentation to said user.

35. A method for analyzing genetic micro array data sets over a data network via a user computer that is executing a thin web client, comprising:
storing the genetic micro array data sets in a common format within a data server;
first transmitting/receiving first transactions over the data network to/from the user computer to configure specific analyses to be conducted on specific genetic micro array data sets;
within an analysis server, executing the specific analyses, said executing yielding results corresponding to each of the specific analyses; and
second transmitting/receiving second transactions over the data network to/from the user computer to provide a user with the results.

36. The method as recited in claim 35, wherein said storing comprises:
uploading the specific genetic micro array data sets over the data network from the user computer.

37. The method as recited in claim 36, wherein said uploading provides some of the specific genetic micro array data sets as data resulting from cDNA chips.

38. The method as recited in claim 36, wherein said uploading provides some of the specific genetic micro array data sets as data resulting from oligonucleotide chips.

39. The method as recited in claim 35, wherein said first transmitting/receiving comprises:
generating commands to transfer configuration templates to the user computer; and
interpreting responses to receive configuration parameters from the user computer;
wherein said generating and said interpreting enable the user to configure the specific analyses.

40. The method as recited in claim 35, wherein said second transmitting/receiving comprises:

generating result presentation templates, wherein said generating enables the user to view the results.

41. The method as recited in claim 40, wherein each of the result presentation templates correspond to each of the specific analyses, and wherein said generating allows the user to simultaneously view the each of the result presentation templates.

42. The method as recited in claim 41, wherein first result presentation templates comprise controls and result presentation areas that allow said user to selectively view correlated cel values for each of a plurality of gene sets, wherein the each of a plurality of gene sets is identified by one of the analyses.

43. The method as recited in claim 42, wherein the result presentation areas distinguish said correlated cel values by color coding.

44. The method as recited in claim 35, wherein said executing comprises:
first performing unsupervised analyses, the unsupervised analyses comprising a plurality of first application program modules, the plurality of first application program modules implementing unsupervised genetic analysis techniques selected from a set that includes Hierarchical Clustering, K-Means Clustering, Principal Component Analysis, and Self-Organizing Map..

45. The method as recited in claim 44, wherein said executing further comprises:
second performing supervised analyses, the supervised analyses comprising a plurality of second application program modules implementing supervised genetic analysis techniques.

46. The method as recited in claim 45, wherein the plurality of first application program modules and the plurality of second application program modules are coded in C/C++.

47. The method as recited in claim 40, wherein said generating comprises:
providing the result presentation templates as Java® applets.

48. The method as recited in claim 35, wherein said storing comprises:
converting data from the genetic micro array data sets to the common format;
imputing particular cel values into particular genetic micro array data sets.

49. The method as recited in claim 35, further comprising:
querying public data sources for information about interesting genes, wherein the interesting genes are designated by the user; and
aggregating the information for presentation to the user.

FIG. 1

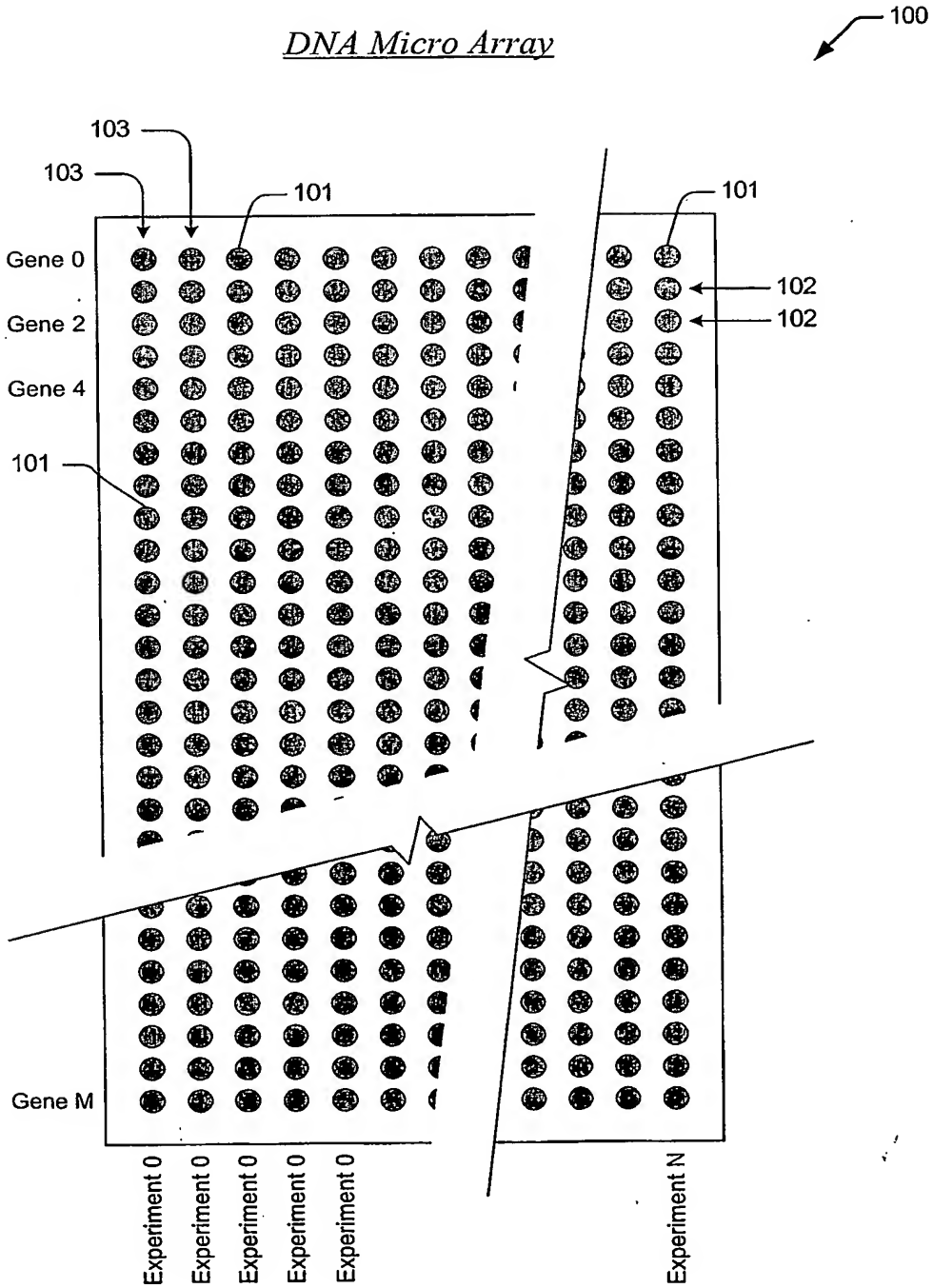


FIG. 2

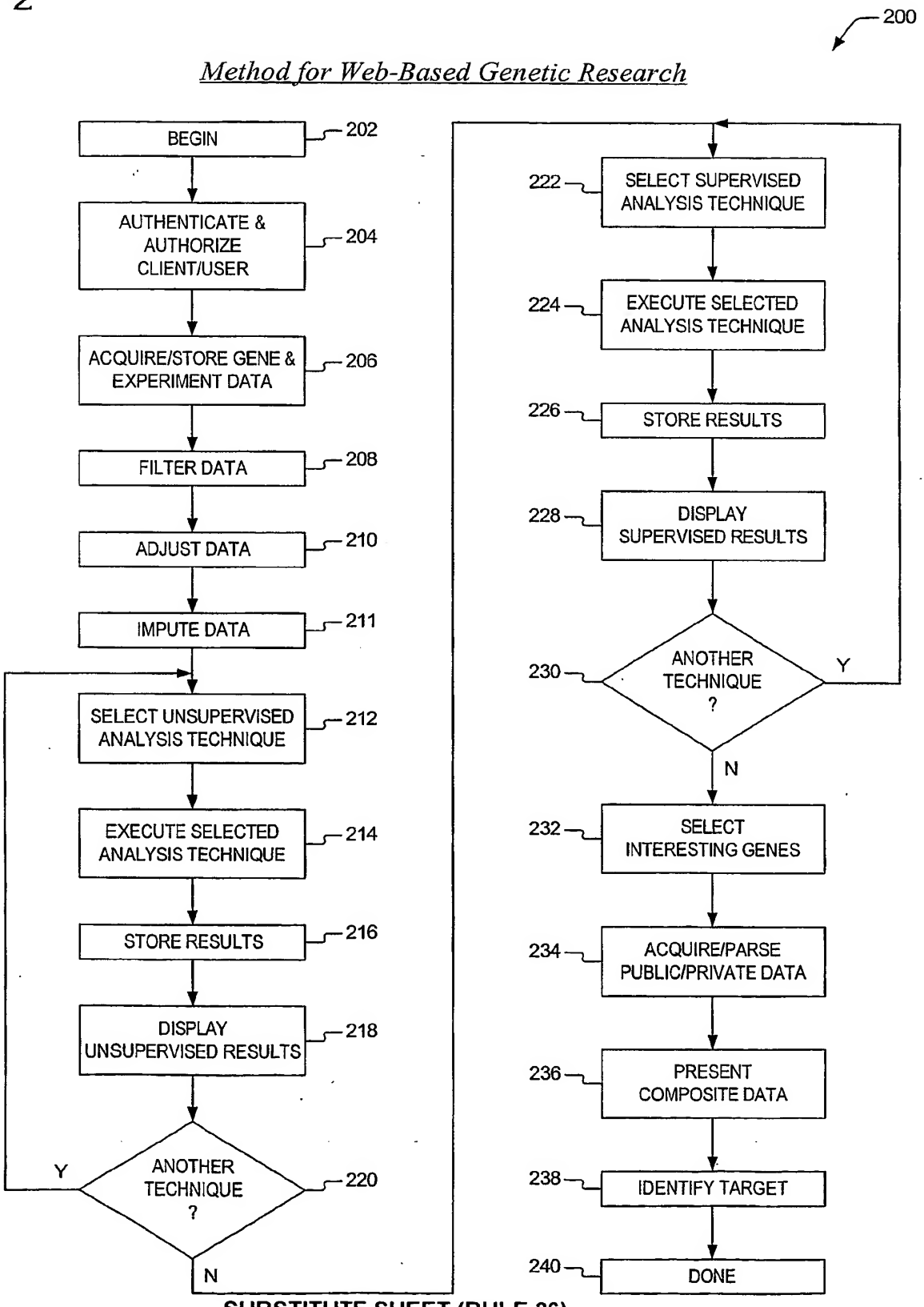


FIG. 3A

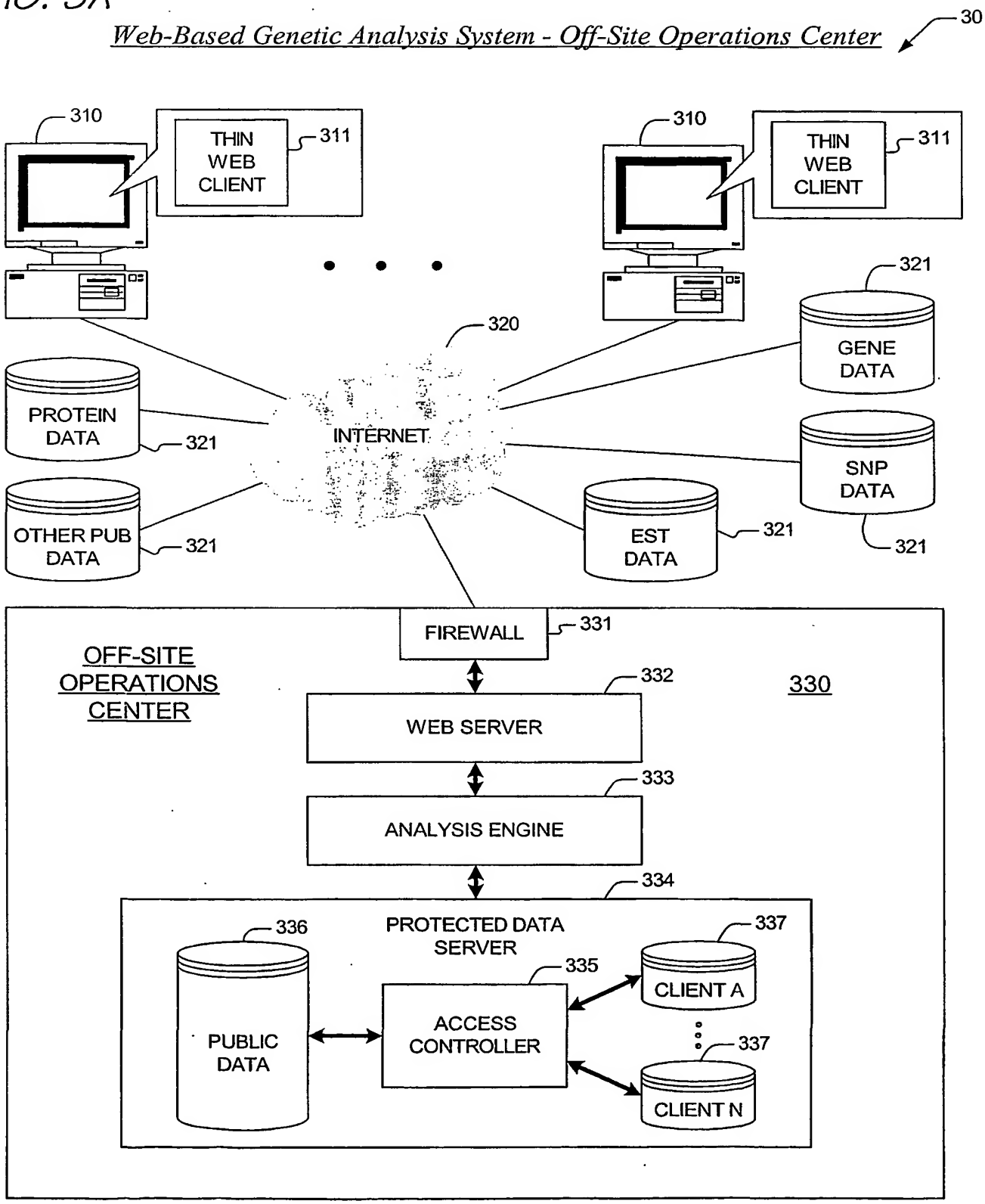


FIG. 3B

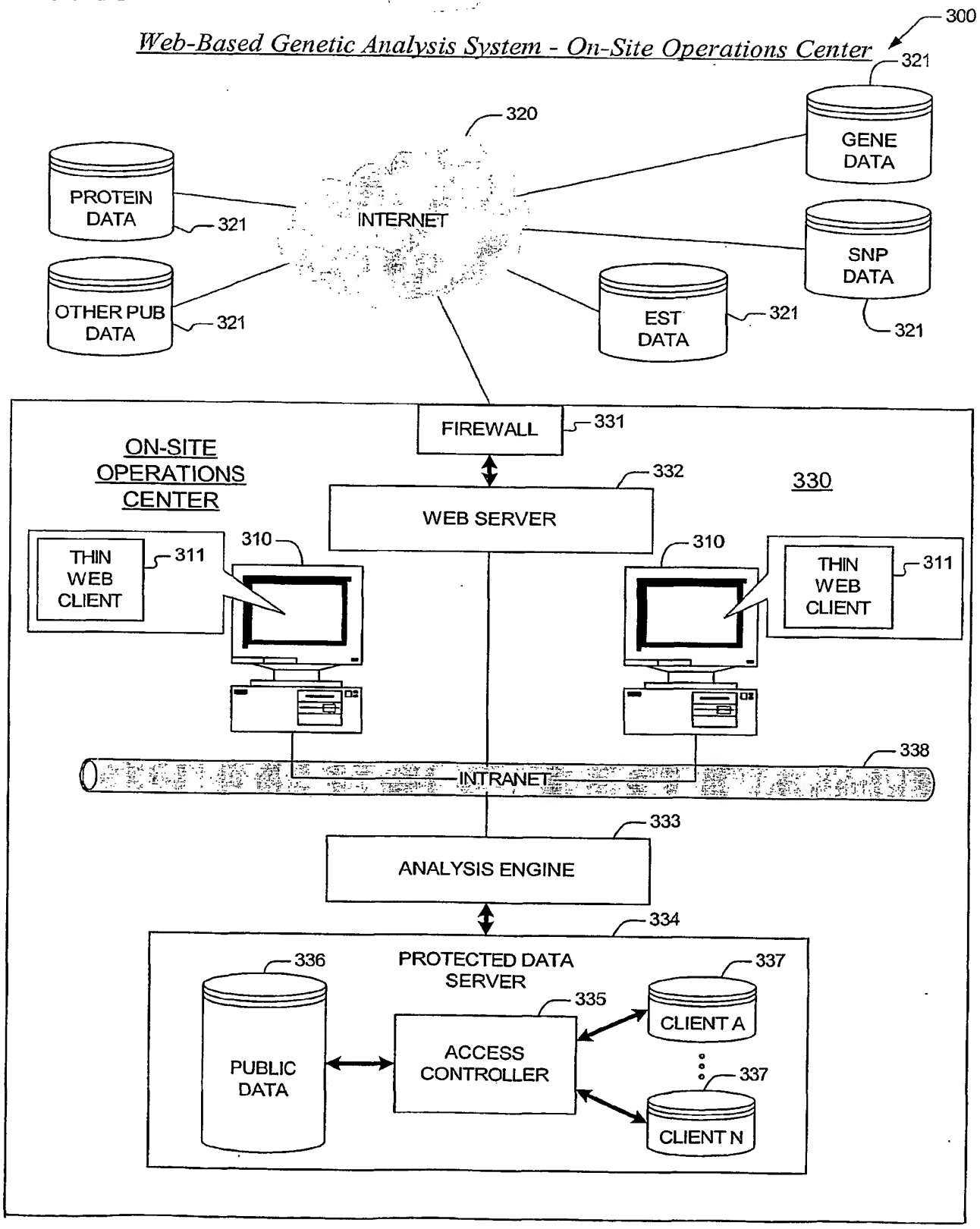


FIG. 4

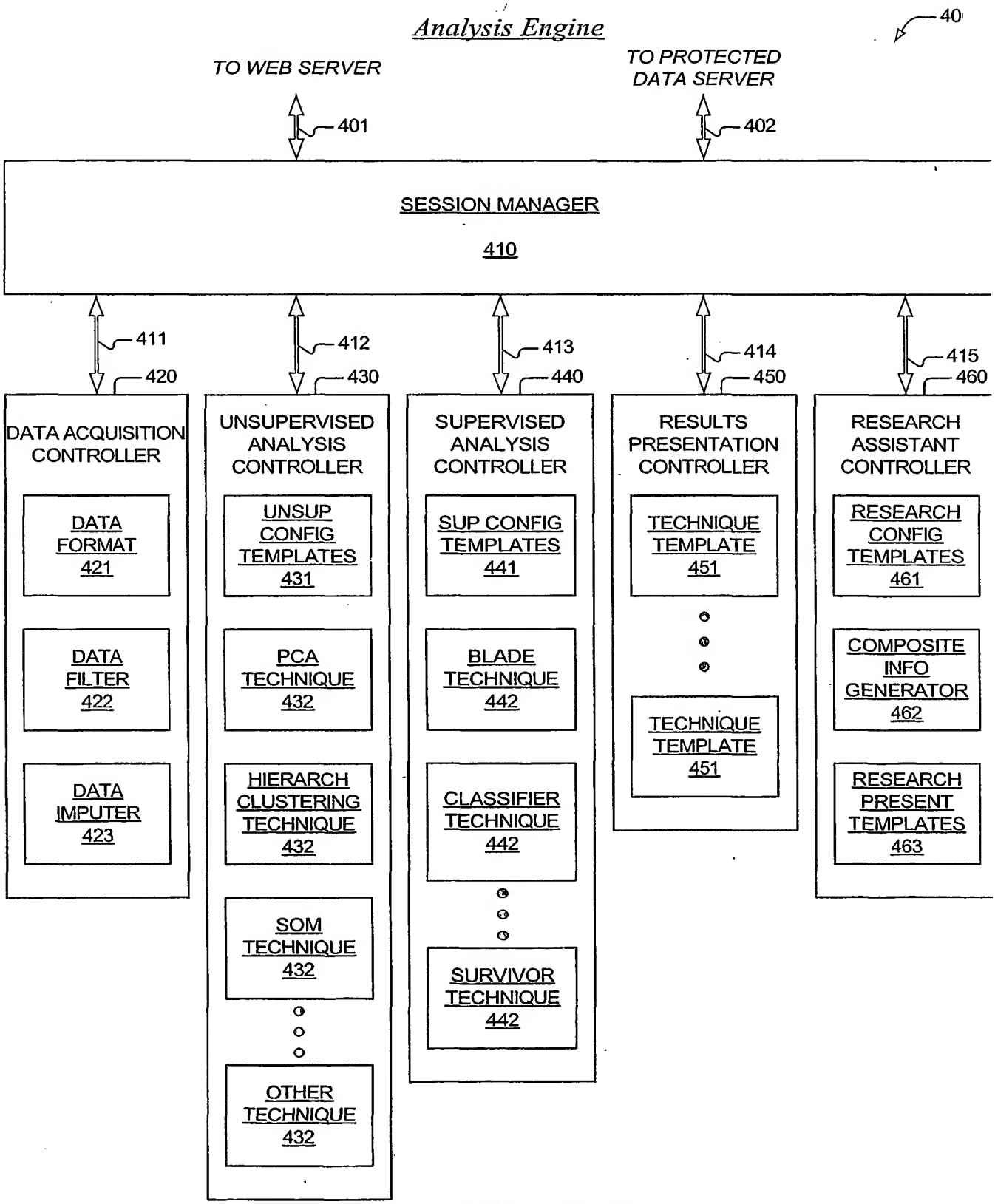


FIG. 5

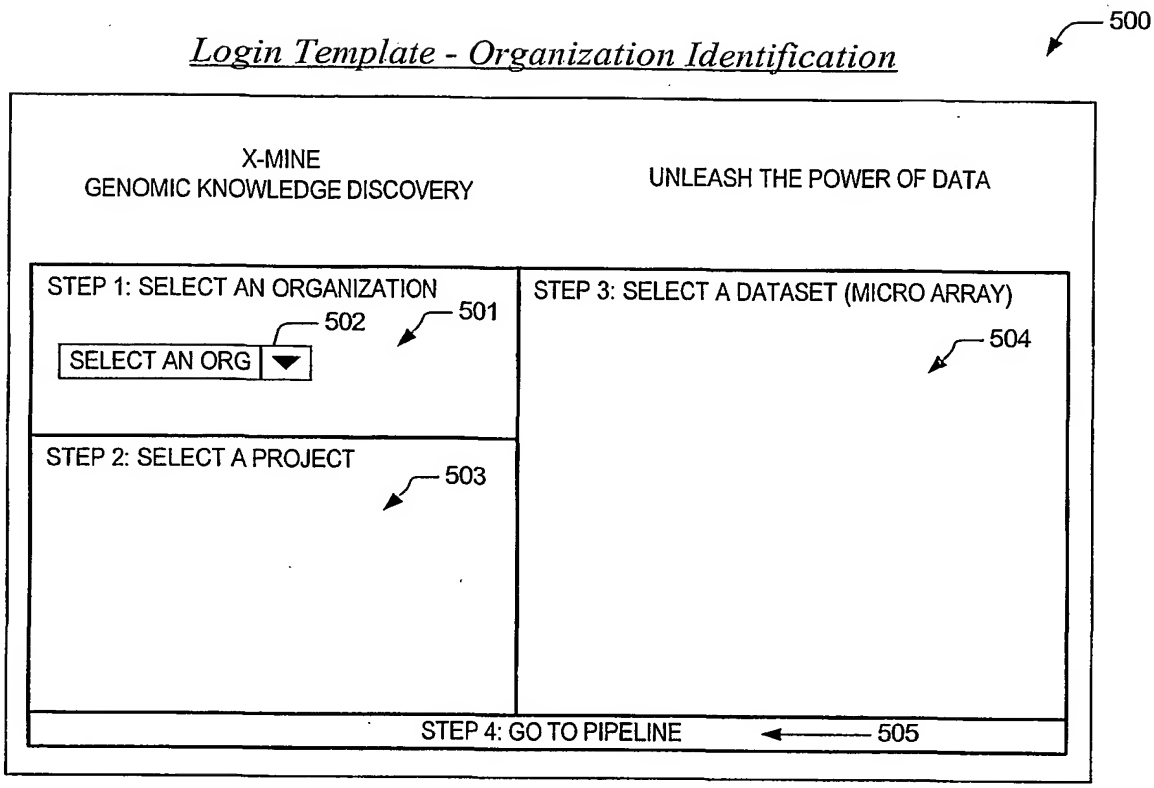


FIG. 6

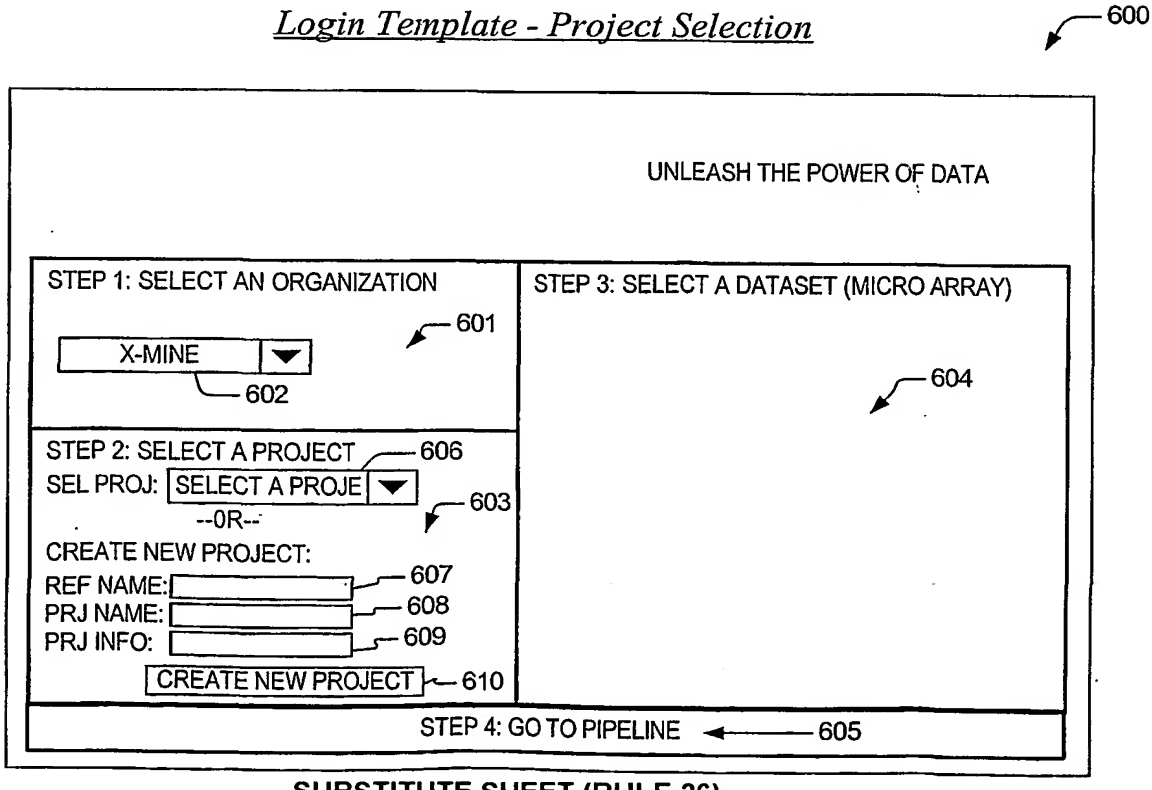


FIG. 7

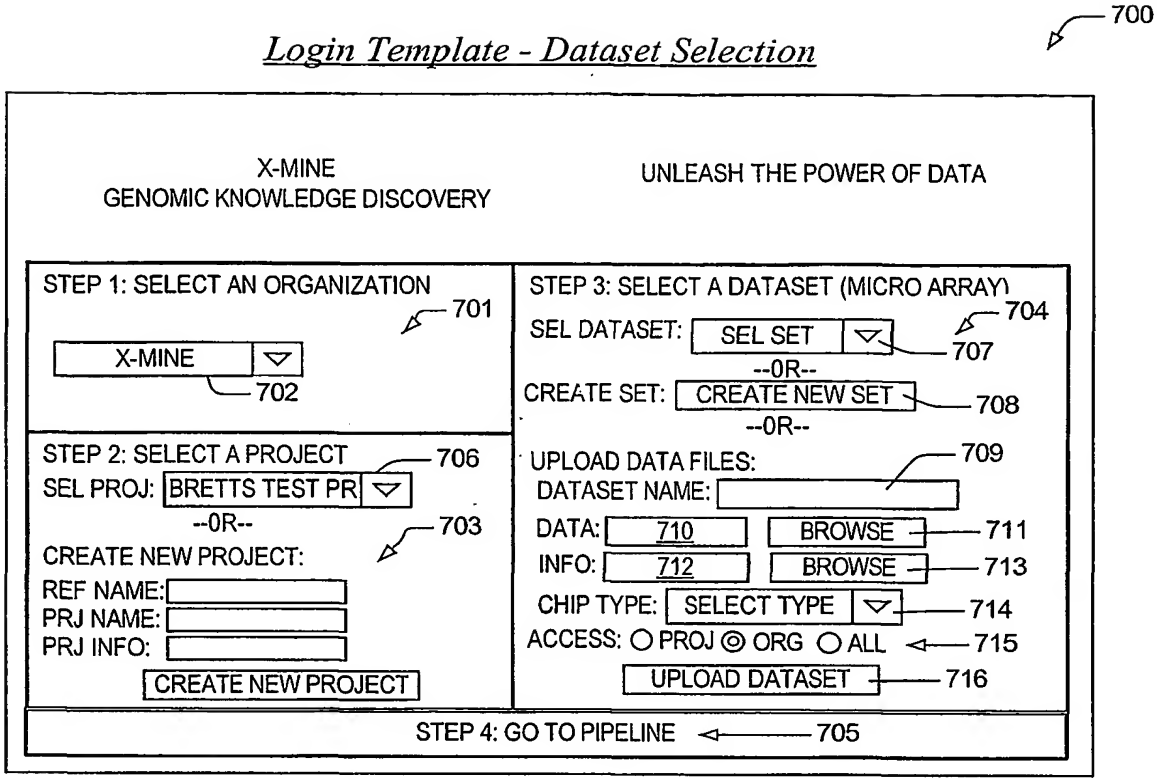


FIG. 8

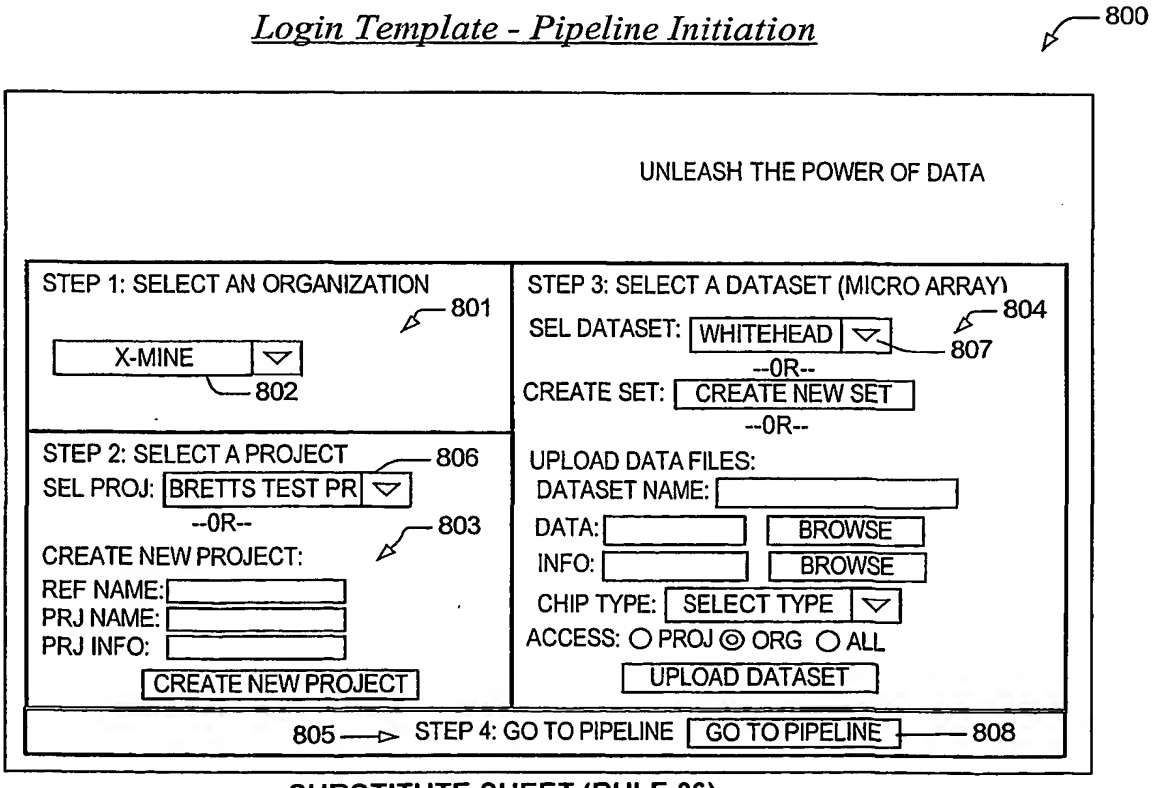


FIG. 9

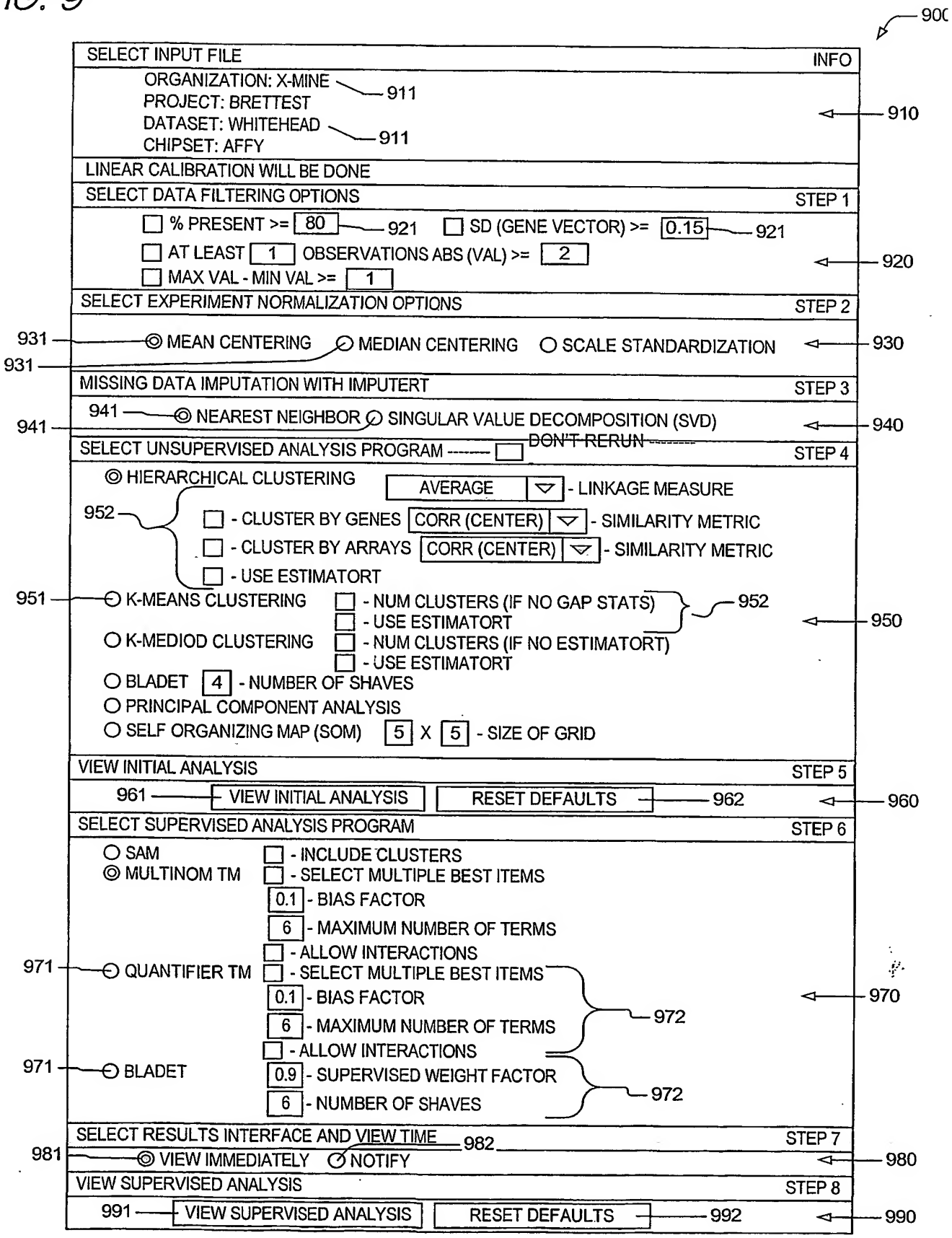


FIG. 14

Web Server Details Featuring Plug-in Application Support